

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-184351

(43)Date of publication of application : 06.07.2001

(51)Int.Cl. G06F 17/28  
G06F 17/22  
G06F 17/30

(21)Application number : 11-370936 (71)Applicant : TOSHIBA CORP

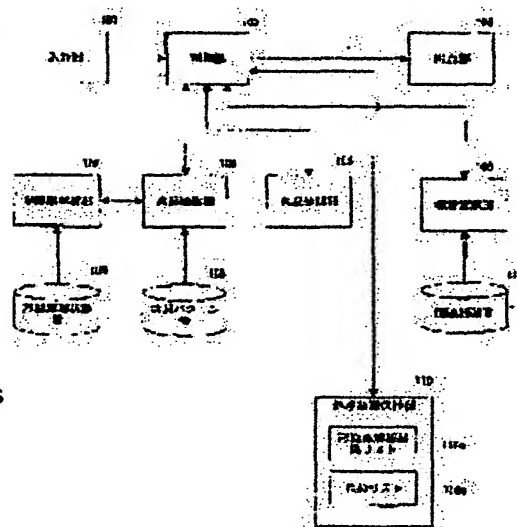
(22)Date of filing : 27.12.1999 (72)Inventor : ICHIMURA YUMI

**(54) DOCUMENT INFORMATION EXTRACTING DEVICE AND DOCUMENT SORTING DEVICE****(57)Abstract:**

**PROBLEM TO BE SOLVED:** To solve a problem that a means for a user to update a dictionary is not provided since the dictionary required for sorting documents is conventionally prepared in advance.

**SOLUTION:** While utilizing a concept definition dictionary in which a concept belonging to each of sorting axes an expression representing this concept are correlated for a plurality of preset plural sorting axes, concepts contained in each of documents are extracted and the document is sorted while using a compound concept created by combining concepts belonging to different sorting axes among the extracted concepts.

Concerning such a system, whether it is necessary to update the dictionary is judged and when updating is required, the list of candidates to be registered in the dictionary is presented so that the document sorting system facilitated in the maintenance of the dictionary can be provided.

**LEGAL STATUS**

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's  
decision of rejection]

[Date of requesting appeal against  
examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2000 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-184351

(P2001-184351A)

(43) 公開日 平成13年7月6日 (2001.7.6)

(51) Int.Cl.<sup>7</sup>

識別記号

F I

テーマコード\* (参考)

G 0 6 F 17/28

G 0 6 F 15/38

C 5 B 0 0 9

17/22

15/20

5 2 0 L 5 B 0 7 5

17/30

15/40

5 2 2 L 5 B 0 9 1

3 7 0 A

3 7 0 J

審査請求 未請求 請求項の数 6 O L (全 18 頁) 最終頁に続く

(21) 出願番号

特願平11-370936

(22) 出願日

平成11年12月27日 (1999. 12. 27)

(71) 出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72) 発明者 市村 由美

神奈川県川崎市幸区小向東芝町1番地 株式会社東芝研究開発センター内

(74) 代理人 100081732

弁理士 大胡 典夫 (外1名)

Fターム(参考) 5B009 MB22 ME13 ME22 MF03 VA02

5B075 ND03 NR12 PP02 PP03 PQ02

QP03 UU06

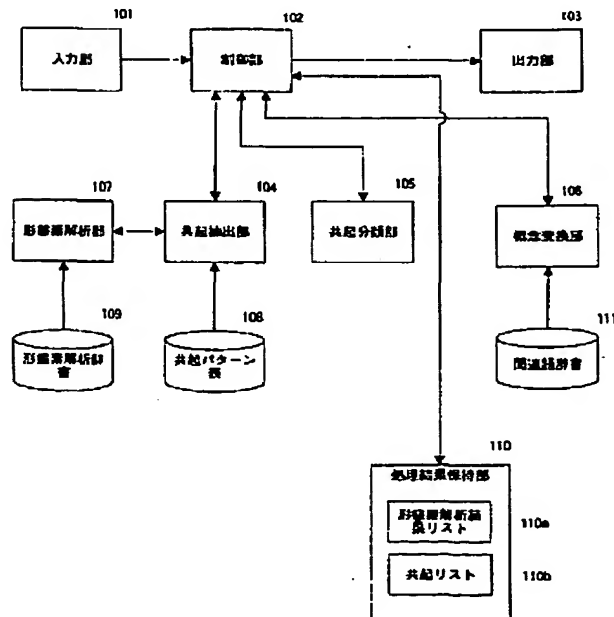
5B091 AA11 AB17 CC05

(54) 【発明の名称】 文書情報抽出装置および文書分類装置

(57) 【要約】

【課題】 従来、文書分類に必要な辞書はあらかじめ用意されているものであり、ユーザが辞書を更新する手段は提供されていなかった。

【解決手段】 あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書を利用して、各文書中に含まれる概念を抽出し、抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分類するシステムにおいて、辞書の更新が必要かどうかを判断し、更新の必要がある場合には辞書登録すべき候補の一覧を提示することにより、辞書のメンテナンスが容易な文書分類システムを提供する。



**【特許請求の範囲】**

【請求項 1】 文書を入力する入力手段と、この入力手段により入力された文書を解析し、あらかじめ登録された共起パターンに適合する共起データを、少なくとも第一の構成要素、第 2 の構成要素、要素間の関係、第 2 の構成要素に付随する付属語列とともに記憶する共起抽出手段と、この共起抽出手段により抽出された共起データに関して、要素間の関係または第 2 の構成要素に付随する付属語列を参照して分類する共起分類手段とを具備することを特徴とする文書情報抽出装置。

【請求項 2】 文書を入力する入力手段と、この入力手段により入力された文書を解析し、あらかじめ登録された共起パターンに適合する共起データを、少なくとも第一の構成要素、第 2 の構成要素、要素間の関係、第 2 の構成要素に付随する付属語列とともに記憶する共起抽出手段と、この共起抽出手段により抽出された共起データに関して、要素間の関係または第 2 の構成要素に付随する付属語列を参照して分類する共起分類手段と、同義語、類義語、シソーラス等の関連語を記憶した関連語辞書と、前記共起抽出手段により抽出された各共起データの構成要素を関連語辞書に記憶される関連語に置き換え、各共起データをより上位の概念へと変換する概念変換手段とを具備することを特徴とする文書情報抽出装置。

【請求項 3】 複数の文書を入力する入力手段と、あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書と、この概念定義辞書を用いて、前記入力手段により入力された各文書中に含まれる概念を抽出する概念抽出手段と、この概念抽出手段により抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせ合わせた複合概念を用いて前記入力された文書を分類する文書分類手段と、この文書分類手段の分類性能を評価する分類性能評価手段と、この分類性能評価手段の評価結果に基づき、前記概念定義辞書への辞書登録候補を自動生成する辞書作成手段と、この辞書作成手段により作成された辞書登録候補を表示する表示手段とを具備したことを特徴とする文書分類装置。

【請求項 4】 前記分類性能評価手段は、各分類に属す文書件数とあらかじめ設定した値とを比較することにより性能評価を行うことを特徴とする請求項 3 記載の文書分類装置。

【請求項 5】 前記分類性能評価手段は、入力された文書件数、いずれかの分類に属す文書件数、および、あらかじめ設定した値とを比較することにより性能評価を行うことを特徴とする請求項 3 記載の文書分類装置。

【請求項 6】 前記辞書作成手段は、入力手段により入力された文書を解析し、あらかじめ登録した共起パターンに適合する共起データを、少なくとも、第 1 の構成要素、第 2 の構成要素、要素間の関係、第 2 の構成要素に

付随する付属語列とともに記憶する共起抽出手段と、共起抽出手段により抽出された共起データに関して、要素間の関係または第 2 の構成要素に付随する付属語列を参照して、各共起データが属す分類軸を推定する分類軸推定手段と、同義語、類義語、シソーラス等の関連語を記憶した関連語辞書と、共起抽出手段により抽出された各共起データの構成要素を関連語辞書に記憶された関連語に置き換え、各共起データをより上位の概念へと変換する概念変換手段とを有し、共起データ、分類軸、上位概念を対応づけて記憶することを特徴とする請求項 3 記載の文書分類装置。

**【発明の詳細な説明】****【0001】**

【発明の属する技術分野】本発明は、文章形式で記述されたデータを自動的に分類する文書分類システムにおいて利用される辞書を自動生成する文書情報抽出方法に関する。

**【0002】**

【従来の技術】さまざまな場所で日々多様な情報が蓄積されているが、そのデータの 8 割以上が文章データであると言われている。このような文章データから情報を引き出すためには、基本的には人間が読むしかないのが現状である。

【0003】たとえば、大量の営業報告書を分析することを考えると、目的に応じて必要なデータだけを形式を決めて記述しておけば、その分析は比較的容易である。しかし、そのような制限を設けると、設定された形式では記述できない情報は捨てられることになる。一方、営業活動における危機警告情報やノウハウ情報などをうまく捉える項目を予め用意することは困難である。そのため、報告書は文章記述を中心に作成されることになり、そのように作成された大量の文書から目的とする情報や知識を読み取るには、大変な労力が必要となる。

【0004】その結果、管理者が業務上の意志決定をするために重要となる営業活動における危機警告情報やノウハウ情報は大量の書類の中に埋もれてしまい、営業報告書が十分に活用されているとはいえないのが現状である。

【0005】上記問題点を解決する手段として、第 57 回情報処理学会全国大会 3-75 および 3-76 にて、テキストマイニング技術によって文書を分析する方法が提案されている。

【0006】この方法では、意図を含んだ情報を抽出するために、モダリティと構文情報を利用している。まず、あらかじめ「要望」「質問」「好評」「問題」といったモダリティに対応する評価ラベルを用意する。そして、構文情報として係り受け関係を想定した 2 語組を抽出し、重要語ラベルと述語・評価ラベルとの組合せパターンを用意して、情報抽出を行う。すなわち、重要語ラベルと呼ぶ名詞表現と評価ラベルを含んだ述語表現との

組合せによって、情報抽出を行っている。

【0007】この方法によると、抽出される概念は、「○が□した」（ただし、□は評価ラベルつき）という複合概念である。つまり、抽出される内容は、「○を□したい」「○を□したくない」「○は○であった」といったような、いわば断片的な情報である。

【0008】一方、業務上の意志決定に有効な情報とは、例えば「プロジェクトの進行状況はどうなっているか?」「あるアクションの結果はどうなっているのか?」「ある事例に遭遇した場合どういう行動をとればよいのか?」といったことであり、そのためにはさらに高度なテキストマイニング技術が求められる。

【0009】上記問題を解決する手段として、当社先願発明である「業務支援システム」（市村、中山）において、アクションと結果を含む複数の分類軸に属す概念とその表現とを定義した概念定義辞書を利用して、各文書中に含まれる概念を抽出し、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分析する方法が提案されている。しかし、前記特許においては、概念定義辞書はあらかじめ用意されているものであり、ユーザがその辞書を更新する手段は提供されていなかった。

【0010】一方、文書から共起データを抽出する方法が、特許第2943447号「テキスト情報抽出装置とテキスト類似照合装置とテキスト検索システムとテキスト情報抽出方法とテキスト類似照合方法、及び、質問解析装置」に開示されている。この方法によると、与えられた文書を解析して共起データを抽出し、さらに抽出した共起データに対して関連語辞書を利用した補完を行う。

【0011】また、抽出された共起データを補完する方法が、特公平5-346938号公報「共起データの補完方式」に開示されている。この方法によると、抽出された共起データに関する統計情報を利用して第1および第2の構成単語間の意味的な距離を算出し、第1の構成単語と第2の構成単語の組合せのうち、もとの構成単語と意味的に近いものを共起データとして補完する。

【0012】上記2つの方法は、文書から共起データを抽出する方法について開示しているが、抽出した共起データをどの分類軸に属す概念であるか推定する方法や、抽出した共起データ同士に共通する概念見出しを付与する方法については述べていない。

【0013】

【発明が解決しようとする課題】以上述べたように、従来、文書分類に必要な辞書はあらかじめ用意されているものであり、ユーザが辞書を更新する手段は提供されていなかった。

【0014】本発明は、この点に鑑み、あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書を利用して、各文書中に含まれる概念を抽出し、抽出された概念

のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分類するシステムにおいて、辞書の更新が必要かどうかを判断し、更新の必要がある場合には辞書登録すべき候補の一覧を提示することにより、辞書のメンテナンスが容易な文書分類システムを提供することを目的とする。

【0015】

【課題を解決するための手段】そこで、本願発明の文書情報抽出方法は、文書を入力する入力部と、前記入力部により入力された文書を解析し、あらかじめ登録した共起パターンに適合する表現を、少なくとも、第1の構成要素、第2の構成要素、要素間の関係、第2の構成要素に付随する付属語列とともに記憶する共起抽出部と、前記共起抽出部により抽出された共起データに関して、要素間の関係または第2の構成要素に付随する付属語列を参照して分類する共起分類部、とを具備することを特徴とする（請求項1）。

【0016】このような構成によれば、与えられた文書から分類軸に関するとともに、共起データを抽出することができる。

【0017】このとき、同義語、類義語、シソーラス等の関連語を記憶した関連語辞書を用意し、抽出された各共起データの構成要素を関連語辞書に記憶された関連語に置き換え、各共起データをより上位の概念へと変換し、共起データ、分類軸、変換された上位概念とを対応づけて記憶することで、各共起データの上位概念を抽出することができる（請求項2）。

【0018】また、本発明の文書分類システムは、複数の文書を入力する入力手段と、あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書と、前記概念定義辞書を用いて、前記入力手段により入力された各文書中に含まれる概念を抽出する概念抽出手段と、前記概念抽出手段により抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて前記入力された文書を分類する文書分類手段と、前記文書分類手段の分類性能を評価する分類性能評価手段と、前記分類性能評価手段の評価結果に基づき、前記概念定義辞書への辞書登録候補を自動生成する辞書作成手段と、前記辞書作成手段により作成された辞書登録候補を表示する表示手段、とを具備することを特徴とする（請求項3）。

【0019】このような構成によれば、分類性能を評価することで、辞書の更新が必要かどうかを判断し、更新の必要がある場合には辞書登録すべき候補の一覧を提示することができる。

【0020】このとき、各分類に属す文書件数とあらかじめ設定した値とを比較したり（請求項4）、入力された文書件数といずれかの分類に属す文書件数とあらかじめ設定した値とを比較することで（請求項5）、分類性能評価を行うことができる。

【0021】また、上記辞書作成手段は、入力手段により入力された文書を解析し、あらかじめ登録した共起パターンに適合する表現を、少なくとも、第1の構成要素、第2の構成要素、要素間の関係、第2の構成要素に付随する付属語列とともに記憶する共起抽出手段と、共起抽出手段により抽出された共起データに関して、要素間の関係または第2の構成要素に付随する付属語列を参照して、各共起データが属す分類軸を推定する分類軸推定手段と、同義語、類義語、シソーラス等の関連語を記憶した関連語辞書と、共起抽出手段により抽出された各共起データの構成要素を関連語辞書に記憶された関連語に置き換え、各共起データをより上位の概念へと変換する概念変換手段とを有し、共起データ、分類軸、上位概念を対応づけて記憶することを特徴とする（請求項6）。

【0022】このような構成によれば、分類軸や概念見出しとともに、辞書登録すべき候補の一覧を提示することができる。

【0023】

【発明の実施の形態】〔第一の実施形態〕以下、図面を参照して本発明に係る第一の実施形態を説明する。図1は本発明の第一の実施形態に係る文書分類システムの構成を示すブロック図である。なお、本実施形態における文書分類システムは、例えば磁気ディスク等の記録媒体に記録されたプログラムを読み込み、このプログラムによって動作が制御されるコンピュータによって実現される。

【0024】図1において、入力手段としての入力部101は、処理対象となる文書を例えばメモリや磁気ディスク、光ディスクなどから取り込む。

【0025】制御部102は入力部101が取り込んだ情報を受け取り解析した後、各処理部へ必要な情報を送る。各処理部での処理結果は再び制御部102に返され、必要な情報は出力部103を介して出力され、例えばディスプレイに表示される。

【0026】処理結果保持部110は、各処理部の結果を一時的に保持するための記憶領域であり、例えばRAMや磁気ディスクなどからなる。この処理結果保持部110には、形態素解析結果リスト110a、共起リスト110bが設けられている。各リストに記憶される情報については、各処理部の説明の中で詳述する。

【0027】まず、制御部102は、入力部101から取り込まれた文書を共起抽出部104に送る。共起抽出部104は、形態素解析部107を起動する。

【0028】形態素解析部107は形態素解析辞書109を参照しながら、形態素解析を行い、その結果を処理結果保持部110に設けられた形態素解析結果リスト110aに記憶する。形態素解析部107の処理動作については広く公知であるので、説明を省略する。

【0029】つぎに、共起抽出部104は、形態素解析

部107から受け取った形態素解析結果リスト110aから、共起パターン表108に記憶されるパターンと一致する表現を抽出し、その結果を処理結果保持部110に設けられた共起リスト110bに記憶する。

【0030】つぎに、制御部102は、共起抽出部104の処理結果である共起リスト110bを共起分類部105に送る。共起分類部105は、共起リスト110bに記憶される情報のうち、要素間の関係あるいは第2の構成要素に付随する付属語列を参照して、各共起データが属す分類軸を推定し、その結果を共起リスト110bに記憶する。

【0031】つぎに、制御部102は、共起分類部105の処理結果である共起リスト110bを概念変換部106に送る。概念変換部106は、共起リスト110bに記憶される各共起データの構成要素を関連語辞書111に記憶される関連語に置き換え、各共起データをより上位の概念へと変換して概念見出しを作成し、その結果を共起リスト110bに記憶する。

【0032】以上が本システムの概要である。次に、各処理部の詳細についてフローチャートを用いて説明する。

（a）共起抽出部104の処理動作

共起抽出部104の処理動作について説明する。図2は共起抽出部104の処理動作を示すフローチャートである。

【0033】まず、ステップS201で共起リスト110bを初期化し、ステップS202に進む。ステップS202で入力文書を読み込み、ステップS203に進む。ステップS203で、処理中の文番号を示す変数*i*に初期値1をセットし、ステップS204に進む。

【0034】ステップS204で、未処理の文が残っているかどうか判定する。残っている場合はステップS205に進み、残っていない場合は、処理を終了する。

【0035】ステップS205で、*i*番目の文を形態素解析部107に送り、その処理結果を受け取り、ステップS206に進む。

【0036】ステップS206で、処理中の文節番号を示す変数*j*に初期値1をセットし、ステップS207に進む。

【0037】ステップS207で*j*+1<=文節数であるかどうか判定する。*j*+1が文節数に等しいか小さい場合にはステップS208に進み、そうでない場合はステップS213に進む。

【0038】ステップS208で、文節*j*、文節*j*+1に一致する共起パターンが共起パターン表108にあるかどうか判定する。一致するパターンがある場合にはステップS209に進み、ない場合にはステップS212に進む。

【0039】ステップS209で、文節*j*、文節*j*+1からなる共起データが共起リスト110bに登録済みかど

うか判定する。すでに登録されている場合はステップS211に進み、まだ登録されていない場合はステップS210に進む。

【0040】ステップS210で、文節j、文節j+1からなる共起データを共起リスト110bに登録し、ステップS211に進む。ステップS211で、該当する共起データの頻度に1を加算し、ステップS212に進む。

【0041】ステップS212でjを1インクリメントし、ステップS207に戻る。ステップS207からステップS212までの繰り返しにより、i番目の文に存在する共起データの抽出を終了すると、ステップS213に進む。ステップS213で、iを1インクリメントし、ステップS204に戻る。ステップS204からステップS213の繰り返しにより、すべての文に存在する共起データの抽出を終了する。

【0042】図3に共起パターン表108に記憶される情報の例を示す。共起パターン表108は第1要素、第1要素と第2要素の関係、第2要素から構成されている。第1要素と第2要素には品詞情報が、関係には付属語情報が格納されている。付属語を伴わない関係の場合には、関係の欄は空になる。

【0043】図4に共起リスト110bに記憶される情報の例を示す。共起リスト110bは第1要素、第1要素と第2要素の関係、第2要素、第2要素に付随する付属語列、頻度、分類軸、概念見出しから構成されている。共起抽出部104の処理が終了した段階では、図4(a)に示すように、分類軸と概念見出しの欄は空である。

【0044】このようにして、共起抽出部105は、形態素解析部107を起動して形態素解析を行い、その結果を共起パターン表108に記憶される各パターンと照合して一致する情報を抽出し、その結果を共起リスト110bに記憶する。

【0045】(b) 共起分類部105の処理動作つぎに、共起分類部105の処理動作について説明する。図5は共起分類部105の処理動作を示すフローチャートである。

【0046】まず、ステップS501で、共起リスト110bを読み込み、ステップS502に進む。ステップS502で、処理中のリスト番号を示す変数iに初期値1をセットし、ステップS503に進む。ステップS503で $i \leq$ リスト件数であるかどうか判定する。iがリスト件数に等しいか小さい場合はステップS504に進み、そうでない場合は処理を終了する。

【0047】ステップS504で、i番目のリストの第2の構成要素が動詞であるかどうか判定する。動詞である場合はステップS505に進み、そうでない場合はステップS508に進む。

【0048】ステップS505で、i番目のリストの要

素間の関係が「を」であるかどうか判定する。「を」である場合はステップS509に進み、そうでない場合はステップS506に進む。

【0049】ステップS506で、i番目のリストの第2の構成要素に付随する付属語列が「ている」であるかどうか判定する。「ている」である場合はステップS510に進み、そうでない場合はステップS507に進む。

【0050】ステップS507で、i番目のリストの第2の構成要素が動詞「ある」であるかどうか判定する。「ある」である場合には、ステップS510に進み、そうでない場合はステップS509に進む。

【0051】ステップS508で、i番目のリストの第2の構成要素が形容詞または形容動詞であるかどうか判定する。形容詞または形容動詞である場合はステップS510に進み、そうでない場合はステップS511に進む。

【0052】ステップS509で、分類軸を「アクション」とし、ステップS512に進む。ステップS510で、分類軸を「状態」とし、ステップS512に進む。ステップS511で、分類軸を「体言」とし、ステップS512に進む。

【0053】ステップS512で、分類軸を共起リスト110bに追加し、ステップS513に進む。

【0054】ステップS513で、iを1インクリメントし、ステップS503に戻る。ステップS503からステップS513までの繰り返しにより、共起リスト110bに記憶されるすべての共起データについて、分類軸がセットされ、処理を終了する。

【0055】共起分類部105の処理が終了した段階で、共起リスト110bには、図4(b)に示すように分類軸が書き込まれる。

【0056】このようにして、共起分類部105は、共起リスト110bに記憶される情報のうち、要素間の関係あるいは第2の構成要素に付随する付属語列を参照して、各共起データが属す分類軸を推定し、その結果を共起リスト110bに記憶する。

【0057】(c) 概念変換部106の処理動作つぎに、概念変換部106の処理動作について説明する。図6は概念変換部106の処理動作を示すフローチャートである。

【0058】まず、ステップS601で、共起リスト110bを読み込み、ステップS602に進む。ステップS602で、処理中のリスト番号を示す変数iに初期値1をセットし、ステップS603に進む。ステップS603で $i \leq$ リスト件数であるかどうか判定する。iがリスト件数に等しいか小さい場合はステップS604に進み、そうでない場合は処理を終了する。

【0059】ステップS604で、i番目の共起データに関して、第1の構成要素をWORD1に、第2の構成

要素をWORD 2に、要素間の関係をF Z Kに、各々セットし、ステップS 605に進む。

【0060】ステップS 605で、関連語辞書111を参照して、WORD 1の関連語があるかどうか判定する。関連語がある場合は、ステップS 606に進み、ない場合はステップS 607に進む。ステップS 606で、該当する関連語をWORD 1にセットし、ステップS 607に進む。

【0061】ステップS 607で、関連語辞書111を参照して、WORD 2の関連語があるかどうか判定する。関連語がある場合は、ステップS 608に進み、ない場合はステップS 609に進む。

【0062】ステップS 608で、該当する関連語をWORD 2にセットし、ステップS 609に進む。

【0063】ステップS 609で、関連語辞書111を参照して、F Z Kの関連語があるかどうか判定する。関連語がある場合は、ステップS 610に進み、ない場合はステップS 611に進む。ステップS 610で、該当する関連語をF Z Kにセットし、ステップS 611に進む。

【0064】ステップS 611で、WORD 1、F Z K、WORD 2を概念見出しとして共起リスト110bに記憶し、ステップS 612に進む。ステップS 612でiを1インクリメントし、ステップS 603に戻る。ステップS 603からステップS 612までの繰り返しにより、共起リスト110bに記憶されるすべての共起データについて、概念見出しがセットされ、処理を終了する。

【0065】図7に、関係語辞書111に記憶される情報の例を示す。関係語辞書111は、表現とその上位概念から構成される。関連語辞書111には、自立語だけでなく付属語を記憶してもよい。

【0066】概念変換部106の処理が終了した段階で、共起リスト110bには、図4(c)に示すように概念見出しが書き込まれる。

【0067】このようにして、概念変換部106は、共起リスト110bに記憶される各共起データの構成要素を関連語辞書111に記憶される関連語に置き換え、各共起データをより上位の概念へと変換して概念見出しを作成し、その結果を共起リスト110bに記憶する。

【0068】以上のようにして、与えられた文書から、分類軸と概念見出しを伴った共起データを抽出することができる。

【0069】〔第二の実施形態〕次に、本発明に係る第二の実施形態を説明する。図8は本発明の第二の実施形態に係る文書分類システムの構成を示すブロック図である。第二の実施形態は、第一の実施形態で説明した文書情報抽出方法を、文書分類システムに応用した例である。

【0070】図8において、図1と同一部分には同一符

号を付し、異なる部分についてのみ説明する。すなわち、図8の制御部102には、文書分類制御部801、辞書更新制御部802、処理結果保持部110が接続している。文書分類制御部801には、概念抽出部803、文書分類部805、分類性能評価部806、概念定義辞書804が接続している。一方、辞書更新制御部802には、共起抽出部104、共起分類部105、概念変換部106、共起パターン表108、形態素解析部107、形態素解析辞書109、関連語辞書111が接続している。

【0071】辞書更新制御部802以下の処理動作は、図1における制御部102以下の処理動作と同様であるので、説明を省略する。

【0072】処理結果保持部110には、抽出概念リスト110c、分類リスト110d、入力文書数110e、辞書候補リスト110fが追加されている。

【0073】図8において、入力手段としての入力部101は、例えばキーボード、マウス、ペン入力装置などからなり、文字列の入力や選択操作指示などのコマンド入力を取り込む。また、処理対象となる文書を例えばメモリや磁気ディスク、光ディスクなどから取り込む。

【0074】制御部102は入力部101が取り込んだ情報を受け取り解析した後、各処理部へ必要な情報を送る。各処理部での処理結果は再び制御部102に返され、必要な情報は出力部103を介して出力され、例えばディスプレイに表示される。

(d) 制御部102の処理動作

まず、制御部102の処理動作について説明し、全体の処理の流れを述べる。図9は制御部102の処理動作を示すフローチャートである。

【0075】まず、ステップS 901で、文書分類制御部801を起動し、文書分類制御部801に付随する各種処理を実行させ、ステップS 902に進む。ステップS 902で、分類性能評価部806の評価結果を受け取り、ステップS 903に進む。ステップS 903で、評価結果がであるかどうか判定する。である、すなわち分類性能が良いと判定された場合は、ステップS 905に進む。でない、すなわち分類性能が悪いと判定された場合は、ステップS 904に進む。

【0076】ステップS 904で、辞書更新制御部802を起動し、辞書更新制御部802に付随する各種処理を実行させ、ステップS 906に進む。ステップS 906で辞書候補110fリストを受け取り、ステップS 907に進む。ステップS 907で受け取った辞書候補リスト110fを出力部103に送り、処理を終了する。

【0077】一方、ステップS 903からステップS 905に進んだ場合は、ステップS 905で、辞書更新不要のメッセージを出力部103に送り、処理を終了する。

【0078】以上が本システムの概要である。次に、各



処理部の詳細についてフローチャートを用いて説明する。

(e) 概念抽出部803の処理動作

つぎに、概念抽出部803の処理動作について説明する。図10は概念抽出部803の処理動作を示すフローチャートである。

【0079】まず、ステップS1001で、抽出概念リスト110cを初期化し、ステップS1002に進む。ステップS1002で概念定義辞書804を読み込み、ステップS1003に進む。

【0080】この処理により、概念定義辞書804のうち処理に必要な情報が、変数DIC\_ID、DIC\_HYOGEN、DIC\_CNT にセットされる。すなわち、概念定義辞書804のi行目の概念IDがDIC\_ID[i]に、表現がDIC\_HYOGEN[i]に、概念定義辞書804の行数がDIC\_CNT にセットされる。

【0081】ここで、図15に概念定義辞書804に記憶される情報の一例を示す。概念IDを示す記号、分類軸、概念見出し、概念を表す表現から構成される。概念を表す表現は、例えば形態素解析や構文解析の処理結果にアクセスする形で記述してもよい。

【0082】図10の説明に戻り、ステップS1003で処理中の文書を示す変数iに初期値1をセットし、ステップS1004に進む。ステップS1004で、未処理の文書が残っているかどうか判定する。残っている場合はステップS1005に進み、残っていない場合はステップS1012に進む。

【0083】ステップS1005でi番目の文書を読み込み、抽出概念リスト110cに文書IDを追加し、ステップS1006に進む。

【0084】ステップS1006で処理中の概念定義辞書804の行を示す変数jに初期値1をセットし、ステップS1007に進む。ステップS1007でj<=DIC\_CNT であるかどうか判定する。jがDIC\_CNT に等しいか小さい場合は、ステップS1008に進む。そうでない場合はステップS1011に進む。

【0085】ステップS1008で、i番目の文書がDIC\_HYOGEN[j]に適合する部分文字列を含んでいるかどうか判定する。適合する部分文字列を含んでいる場合はステップS1009に進み、そうでない場合はステップS1010に進む。ステップS1009で抽出概念リスト110cにDIC\_ID[j]を追加する。

【0086】ステップS1010でjを1インクリメントし、ステップS1007に戻る。ステップS1007からステップS1010の繰り返しにより、概念定義辞書804のすべての行に対する照合を終了すると、ステップS1011に進む。ステップS1011でiを1インクリメントし、次の文書の処理に移る。

【0087】ステップS1012で、入力文書数110eにi-1を記憶し、処理を終了する。

【0088】図16に抽出概念リスト110cに記憶される情報の例を示す。抽出概念リスト110cは文書ID、抽出された概念IDから構成される。1文書内に複数の概念が抽出された場合には、カンマで区切って概念IDが記述される。

【0089】以上のようにして、概念抽出部803は、概念定義辞書804に記憶される情報を用いて、与えられた文書に含まれる概念を抽出し、その結果を抽出概念リスト110cに記憶する。また、入力された文書の数を入力文書数110eに記憶する。

【0090】(f) 文書分類部805の処理動作

つぎに、文書分類部805の処理動作について説明する。図11は文書分類部805の処理動作を示すフローチャートである。

【0091】まず、ステップS1101で、分類リスト110dを初期化し、ステップS1102に進む。ステップS1102で抽出概念リスト110cと入力文書数110eを読み込み、ステップS1103に進む。

【0092】ステップS1103で概念定義辞書804を読み込み、ステップS1104に進む。この処理により、概念定義辞書804の内容が、変数DIC\_ID、DIC\_JIKU、DIC\_CNT にセットされる。

【0093】ステップS1104で分類項目を作成し、ステップS1105に進む。この処理により、分類すべき項目一覧が、変数bunrui\_id、bunrui\_cntにセットされる。このステップS1104の処理動作については後述する。

【0094】ステップS1105で処理中の文書を示す変数iに初期値1をセットし、ステップS1106に進む。ステップS1106でi<=入力文書数であるかどうか判定する。iが入力文書数に等しいか小さい場合は、ステップS1107に進み、そうでない場合は処理を終了する。

【0095】ステップS1107で処理中の分類項目を示す変数jに初期値1をセットし、ステップS1108に進む。ステップS1108でj<=bunrui\_cntであるかどうか判定する。jがbunrui\_cntに等しいか小さい場合は、ステップS1109に進み、そうでない場合はステップS1114に進む。

【0096】ステップS1109で、i番目の文書はbunrui\_id[j]に含まれる概念のすべてを含むかどうか判定する。bunrui\_id[j]に含まれるすべての概念を含む場合にはステップS1110に進み、そうでない場合はステップS1113に進む。

【0097】ステップS1110で分類リスト110dにbunrui\_id[j]が存在するかどうか判定する。存在しない場合はステップS1111に進み、存在

する場合はステップS1112に進む。ステップS1111で分類リスト110dに**bunrui\_id[j]**を追加し、ステップS1112に進む。ステップS1112で分類リスト110dの**bunrui\_id[j]**の欄に、i番目の文書の文書IDを追加し、ステップS1113に進む。

【0098】ステップS1113でjを1インクリメントし、ステップS1108に戻る。ステップS1108からステップS1113までの繰り返しにより、すべての分類項目との照合を終了すると、ステップS1114に進む。

【0099】ステップS1114でiを1インクリメントし、ステップS1106に戻る。ステップS1106からステップS1114までの繰り返しにより、すべての文書のチェックを終了すると、処理を終了する。

【0100】ここで、ステップS1104の処理動作について説明する。図12はステップS1104の処理動作を示すフローチャートである。

【0101】ステップS1201で処理中の分類項目を示す変数mに初期値1をセットし、ステップS1202に進む。ステップS1202で変数Sに初期値1をセットし、ステップS1203に進む。

【0102】ステップS1203で $S \leq DIC\_CNT$  であるかどうか判定する。SがDIC\_CNT に等しいか小さい場合はステップS1204に進み、そうでない場合はステップS1211に進む。

【0103】ステップS1204で変数tに初期値Sをセットし、ステップS1205に進む。ステップS1205で $t \leq DIC\_CNT$  であるかどうか判定する。tがDIC\_CNT に等しいか小さい場合はステップS1206に進み、そうでない場合はステップS1210に進む。

【0104】ステップS1206で $DIC\_JIKU[s] \neq DIC\_JIKU[t]$  であるかどうか判定する。

【0105】 $DIC\_JIKU[s]$  と $DIC\_JIKU[t]$  が等しくない場合はステップS1207に進み、等しい場合はステップS1209に進む。

【0106】ステップS1207で、変数**bunrui\_id[m]**に**DIC\_ID[s]**と**DIC\_ID[t]**を連結したものをセットし、ステップS1208に進む。

【0107】ステップS1208でmを1インクリメントし、ステップS1209に進む。ステップS1209でtを1インクリメントし、ステップS1205に戻る。ステップS1210でSを1インクリメントし、ステップS1203に戻る。

【0108】ステップS1203からステップS1210までの繰り返しにより、異なる軸に属す概念同士の組合せを作成し、ステップS1211に進む。ステップS

1211で変数**bunrui\_cnt**に分類項目数であるm-1をセットし、処理を終了する。ここでは、2つの軸について組合せを作成しているが、3つ以上の軸について組み合わせを作成するようにしてもよい。

【0109】図17に分類リスト110dに格納される情報の例を示す。分類リスト110dは分類項目、文書ID、文書件数から構成される。1つの分類項目に対して複数の文書が該当する場合には、カンマで区切って文書IDが記述される。

【0110】以上のようにして、文書分類部805は、概念抽出部803の処理結果を受け取り、異なった分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分類し、その結果を分類リスト110dに記憶する。

【0111】(g) 分類性能評価部806の処理動作につき、分類性能評価部806の処理動作について説明する。ここでは、2通りの処理動作について説明する。

【0112】まず、第一の処理動作について説明する。図13は分類性能評価部806の第一の処理動作を示すフローチャートである。

【0113】まず、ステップS1301で、分類リスト110dを読み込み、ステップS1302に進む。ステップS1302で評価結果の初期値をとり、ステップS1303に進む。

【0114】ステップS1303で処理中のリストを示す変数iに初期値1をセットし、ステップS1304に進む。ステップS1304で $i \leq$  リスト件数であるかどうか判定する。iがリスト件数に等しいか小さい場合は、ステップS1305に進み、そうでない場合は処理を終了する。

【0115】ステップS1305でリストのi番目の分類項目に分類される文書件数 $> \alpha$ であるかどうか判定する。文書件数が $\alpha$ より大きい場合はステップS1306に進み、そうでない場合はステップS1307に進む。ここで $\alpha$ はあらかじめ設定しておく値である。

【0116】ステップS1306で評価結果を $\times$ とし、処理を終了する。ステップS1307でiを1インクリメントし、ステップS1304に戻る。ステップS1304からステップS1307までの繰り返しにより、分類リスト110d中に1つでも文書件数が $\alpha$ を超える分類項目があれば、評価結果は $\times$ となり、すべての分類項目について $\alpha$ を超えなければ評価結果は $\circ$ となる。

【0117】次に、第二の処理動作について説明する。図14は分類性能評価部806の第二の処理動作を示すフローチャートである。

【0118】まず、ステップS1401で入力文書数110eと分類リスト110dを読み込み、ステップS1402に進む。

【0119】ステップS1402で、分類リスト110dに記述される文書IDの異なり総数/入力文書数を計算し、その値が $\beta$ より小さいかどうか判定する。

【0120】 $\beta$ より小さい場合はステップS1403に進み、そうでない場合はステップS1404に進む。ここで $\beta$ はあらかじめ設定しておく値である。ステップS1403で評価を $\times$ 、ステップS1404で評価をとし、処理を終了する。

【0121】以上2通りの処理動作に示したように、分類性能評価部806は、分類リスト108bや入力文書数110eを参照して、システムの分類性能を評価する。

【0122】図19に本発明を利用したシステムの出力例を示す。図19(a)は、辞書のチェックを行うかどうかの確認画面である。ここで、ユーザが実行ボタンを押すと、辞書のチェックが実行され、図19(b)または(c)のように辞書チェック結果が表示される。

【0123】図19(b)は辞書更新の必要がないと判断された場合である。図19(c)は辞書更新の必要があると判断された場合である。この場合には、さらに辞書更新を行うかどうかの確認画面が表示される。ここで、ユーザが実行ボタンを押すと、辞書更新が実行され、図19(d)に示すように、登録すべき辞書候補一覧が表示される。ここで、実際に登録したい候補にチェックし、登録ボタンを押すと、辞書に自動登録される。

【0124】なお、本発明は、上記実施形態に限定されず、要旨を変更しない範囲で適宜変更して実施可能である。また、上述した実施形態において記載した方法は、コンピュータに実行させるプログラムとして、例えば、磁気ディスク（フロッピーディスク、ハードディスク等）、光ディスク（CD-ROM、DVD等）、半導体メモリ等の記録媒体に書き込んで各種装置に適用したり、通信媒体により伝送して各種装置に適用することも可能である。本装置を実現するコンピュータは、記録媒体に記録されたプログラムを読み込み、このプログラムによって動作が制御されることにより、上述した処理を実行する。

#### 【0125】

【発明の効果】以上説明したように、上記の実施形態によれば、あらかじめ設定された複数の分類軸に対して、各分類軸に属す概念と該概念を表す表現とを対応づけた概念定義辞書を利用して、各文書中に含まれる概念を抽出し、抽出された概念のうち、異なる分類軸に属す概念同士を組み合わせた複合概念を用いて文書を分類するシステムにおいて、辞書の更新が必要かどうかを判断し、更新の必要がある場合には辞書登録すべき候補の一覧を提示することにより、辞書のメンテナンスが容易な文書分類システムを提供することができる。

#### 【図面の簡単な説明】

【図1】本発明の第一の実施形態に係る文書情報抽出方法の概略構成を示すブロック図。

【図2】同実施形態における共起抽出部104の処理動

作を示すフローチャート。

【図3】共起パターン表108に記憶される情報の一例。

【図4】共起リスト110bに記憶される情報の一例。

【図5】同実施形態における共起分類部105の処理動作を示すフローチャート。

【図6】同実施形態における概念変換部106の処理動作を示すフローチャート。

【図7】関連語辞書111に記憶される情報の一例。

【図8】本発明の第二の実施形態に係る文書分類システムの概略構成を示すブロック図。

【図9】同実施形態における制御部102の処理動作を示すフローチャート。

【図10】同実施形態における概念抽出部803の処理動作を示すフローチャート。

【図11】同実施形態における文書分類部805の処理動作を示すフローチャート。

【図12】図11におけるステップS1104の処理動作を示すフローチャート。

【図13】同実施形態における分類性能評価部806の第一の処理動作を示すフローチャート。

【図14】同実施形態における分類性能評価部806の第二の処理動作を示すフローチャート。

【図15】概念定義辞書804に記憶される情報の一例。

【図16】抽出概念リスト110cに記憶される情報の一例。

【図17】分類リスト110dに記憶される情報の一例。

【図18】辞書候補リスト108dに記憶される情報の一例。

【図19】出力例。

【符号の説明】

101…入力部

102…制御部

103…出力部

104…共起抽出部

105…共起分類部

106…概念変換部

107…形態素解析部

108…共起パターン表

109…形態素解析辞書

110…処理結果保持部

801…文書分類制御部

802…辞書更新制御部

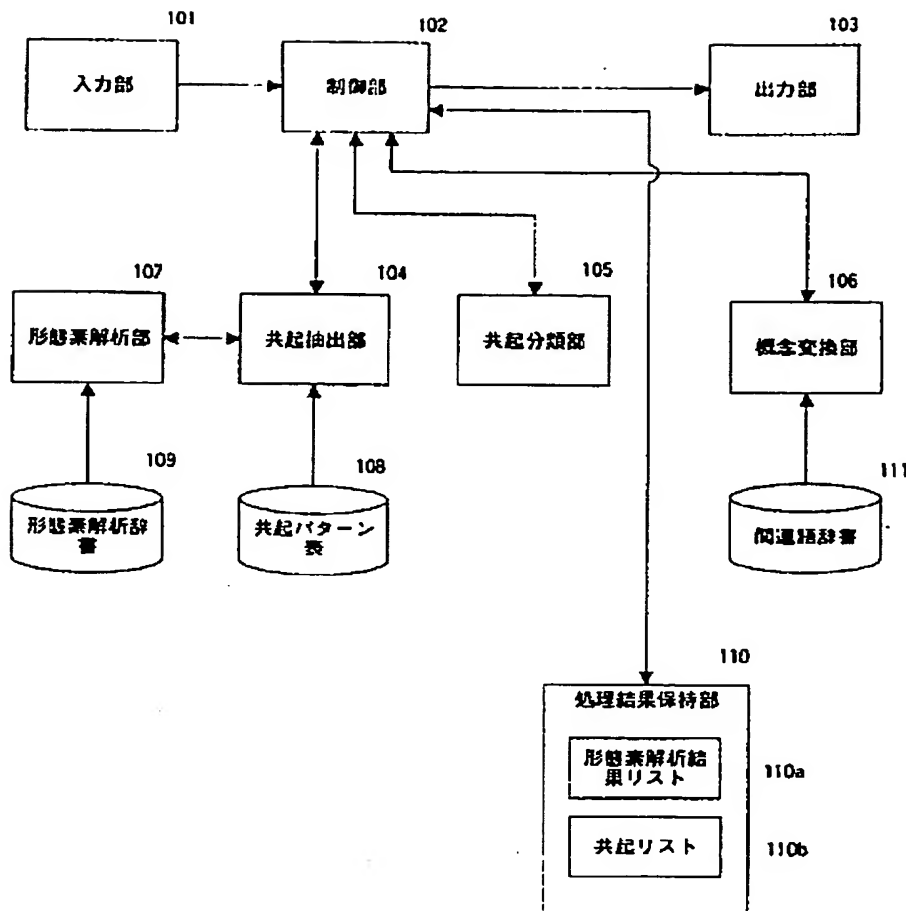
803…概念抽出部

804…概念定義辞書

805…文書分類部

806…分類性能評価部

【図1】



【図7】

表現	上位概念
良くない	悪い
売上げ	売行き

【図17】

分類項目	文書ID	文書件数
A001+S001	#002, .....	100
A001+S002	#001, .....	200
A002+S003	#001, .....	35

【図3】

第1要素	関係	第2要素
名詞	が、は	動詞
名詞	を、も	動詞
名詞	が、は	形容詞
名詞	が、は	形容動詞

【図15】

概念ID	分類軸	概念見出し	表現
A001	アクション	サンプルをつける	サンプルラッピング
S001	状態	売行きが悪い	売れていない
S001	状態	売行きが悪い	動きが悪い

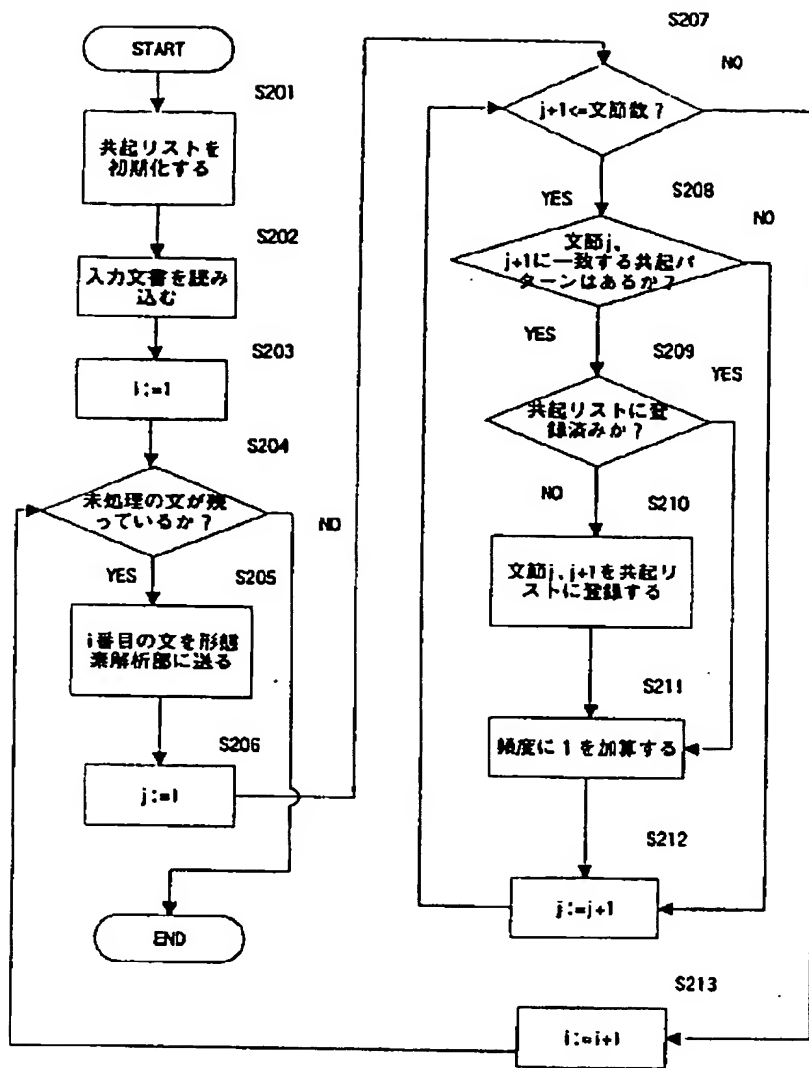
【図16】

文書ID	概念ID
#001	A001, A002, R003, .....
#002	A001, R002, .....
#003	S005, S002, .....

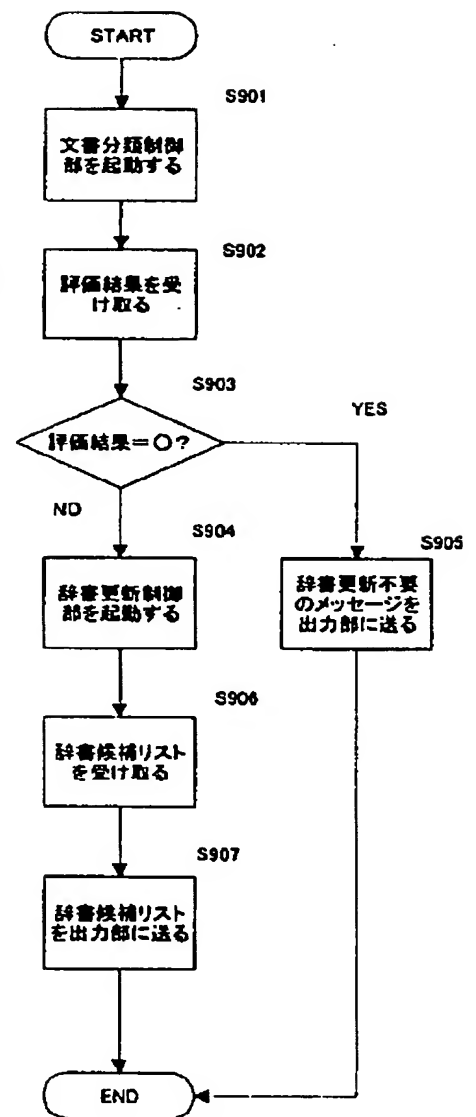
【図18】

概念ID	分類軸	概念見出し	表現
NS001	状態	売行きが悪い	売行きが悪い
NA001	アクション	サンプルを付けた	サンプルを付けた
NS001	状態	売行きが悪い	売上げが良くない

【図 2】



【図 9】



【図 4】

(a)	第1要素	関係	第2要素	付属語列	頻度	分類軸	概念見出し
	売行き	が	悪い		10		
	サンプル	を	付け	た	5		
	売上げ	が	良く	ない	20		

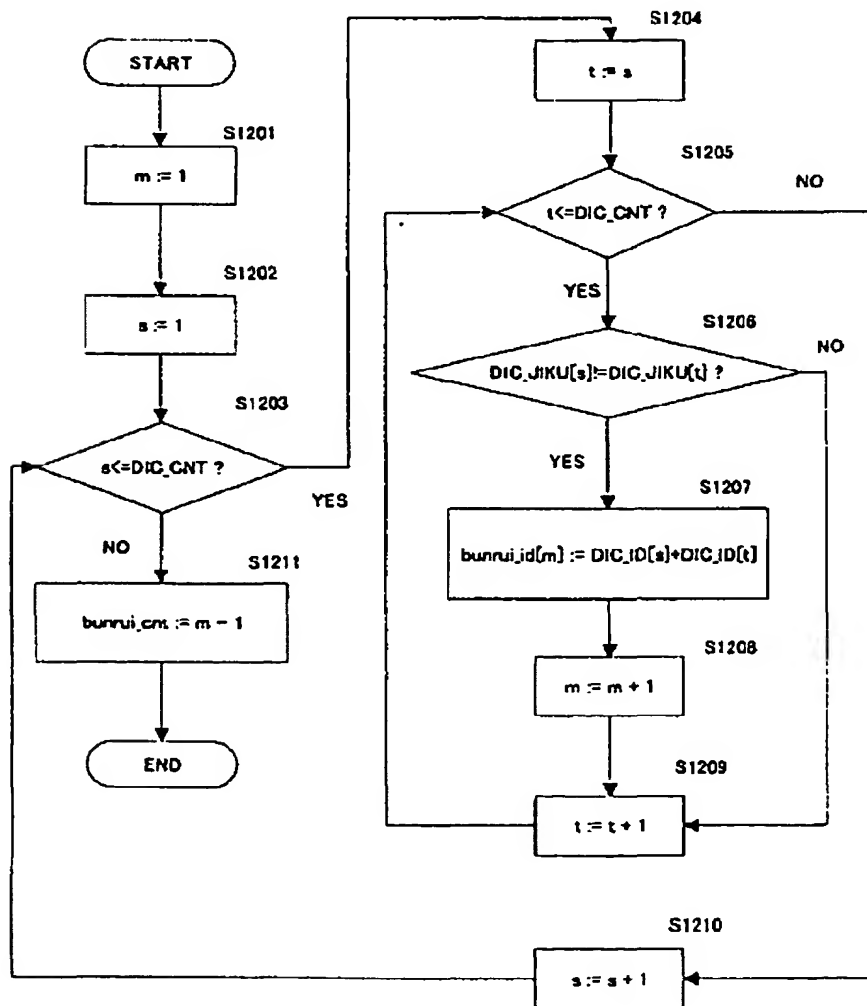
  

(b)	第1要素	関係	第2要素	付属語列	頻度	分類軸	概念見出し
	売行き	が	悪い		10	状態	
	サンプル	を	付け	た	5	アクション	
	売上げ	が	良く	ない	20	状態	

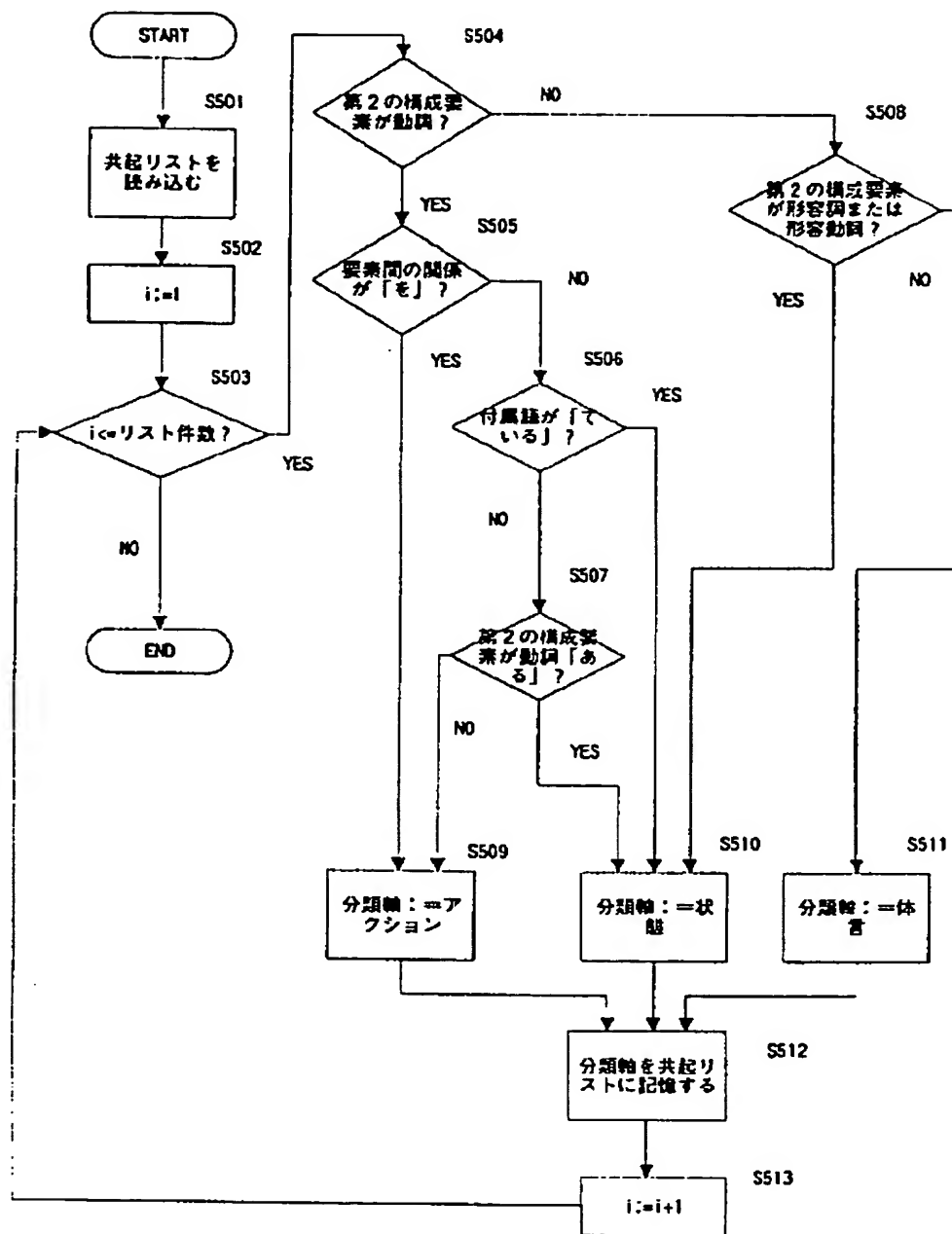
  

(c)	第1要素	関係	第2要素	付属語列	頻度	分類軸	概念見出し
	売行き	が	悪い		10	状態	売行きが悪い
	サンプル	を	付け	た	5	アクション	サンプルを付けた
	売上げ	が	良く	ない	20	状態	売行きが悪い

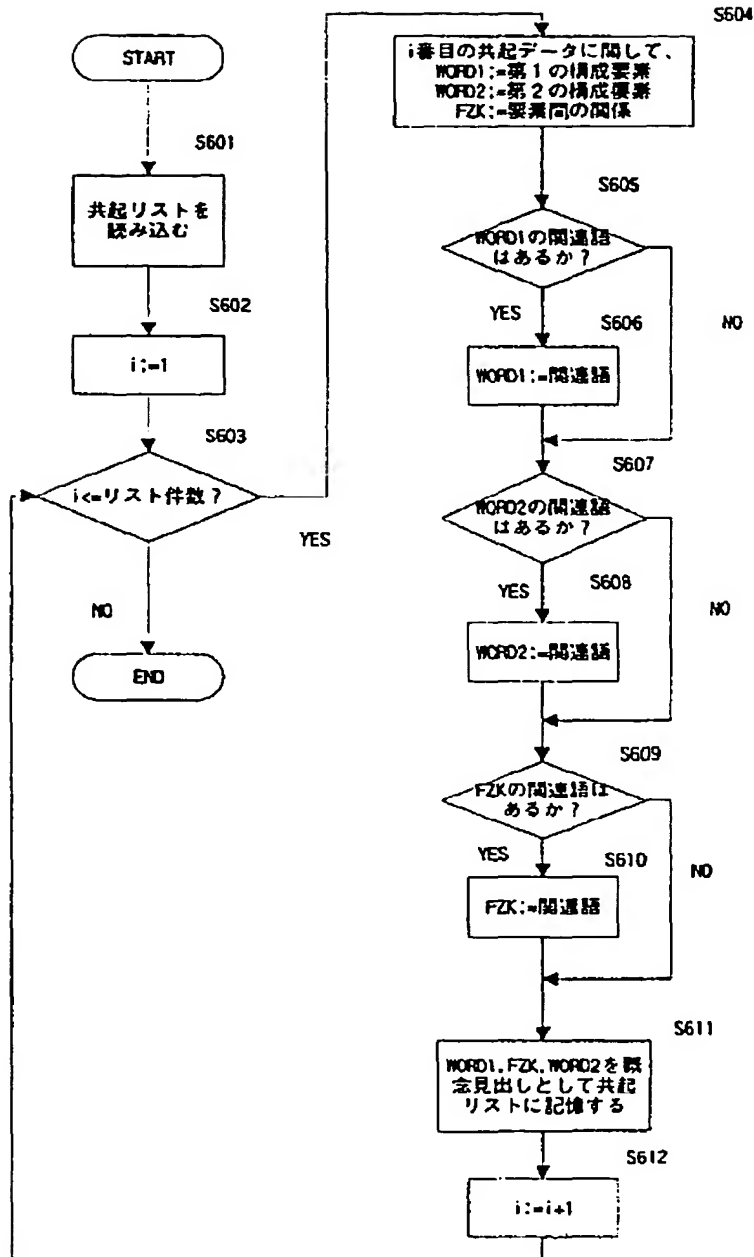
【図 12】



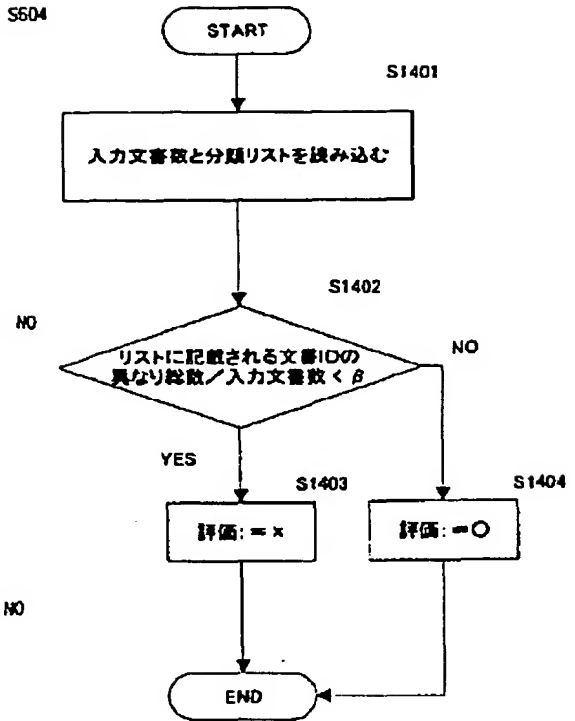
【図5】



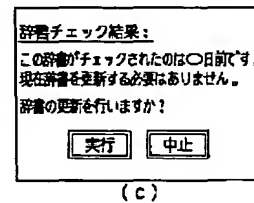
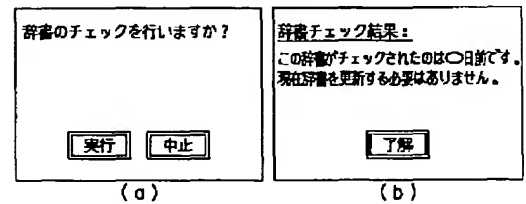
【図6】



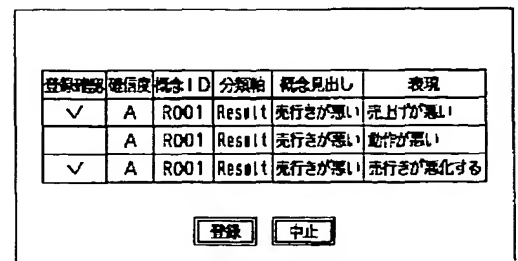
【図14】



【図19】



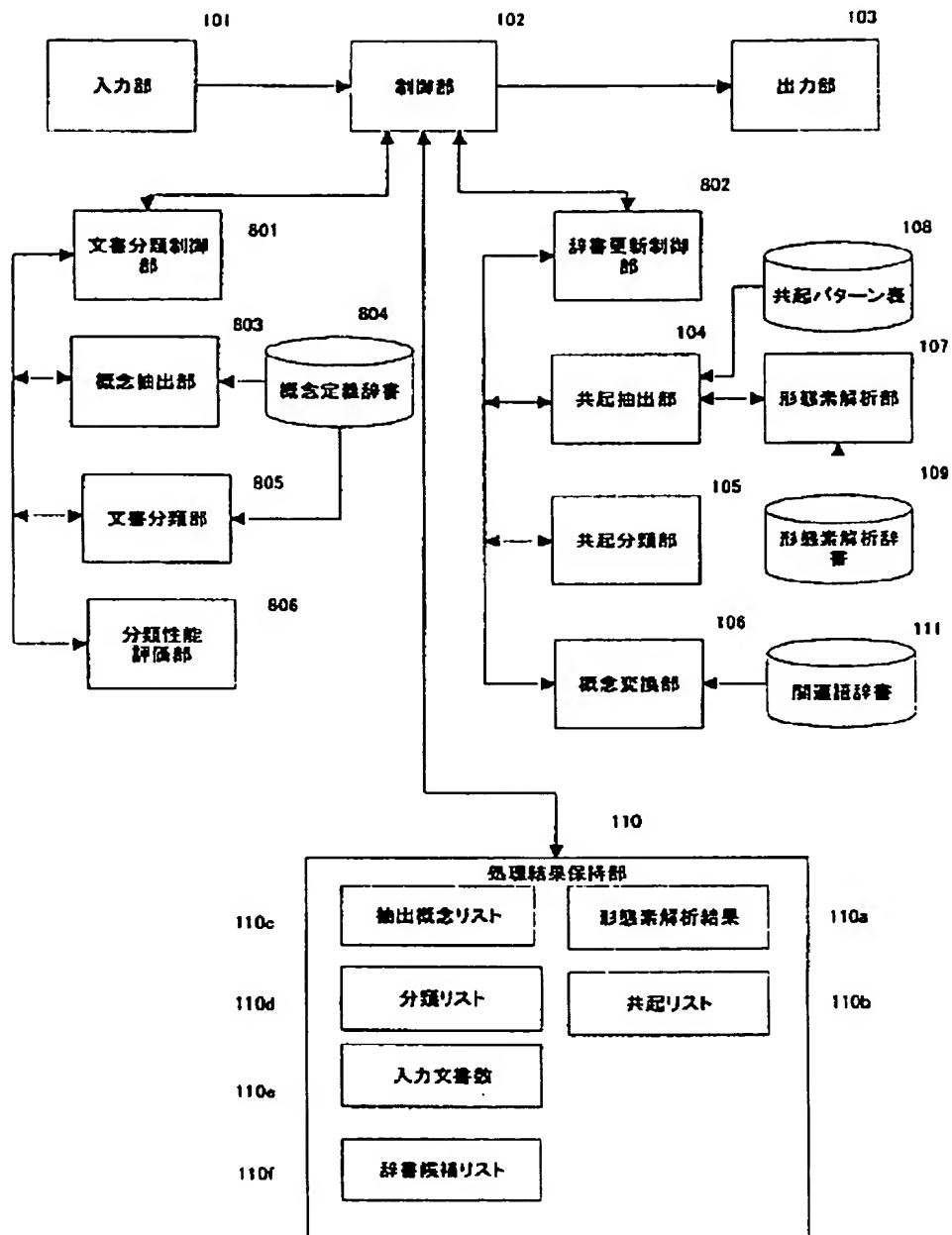
(c)



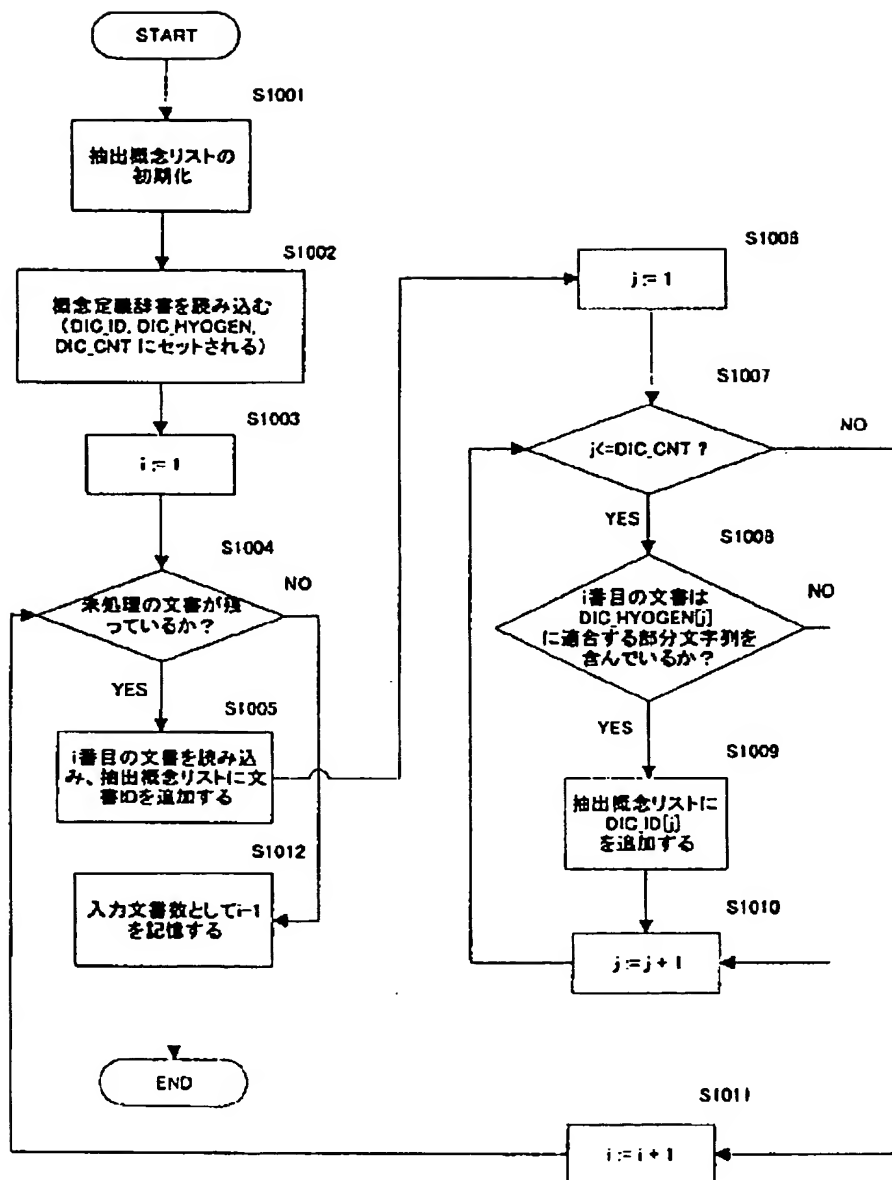
(d)



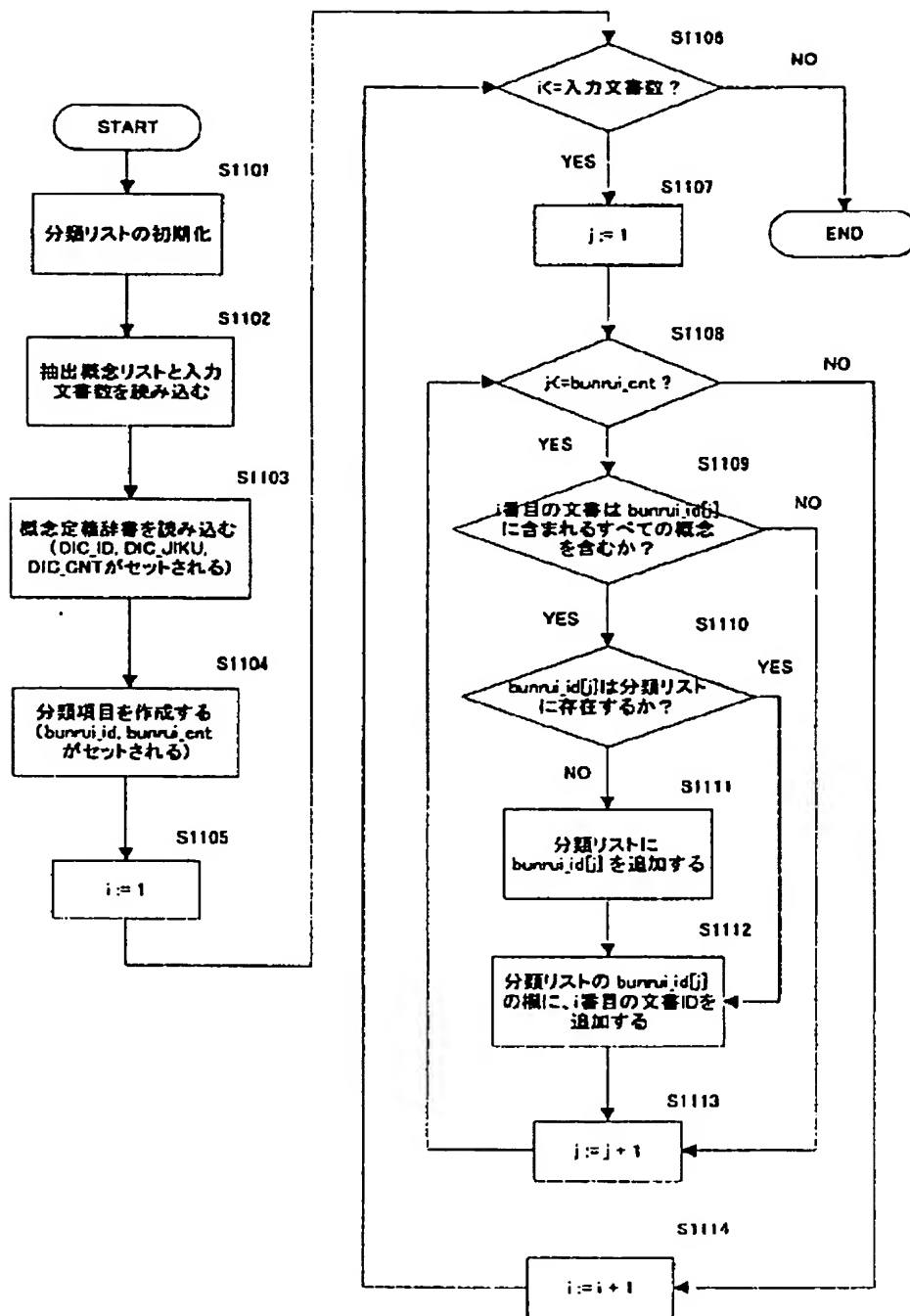
【図 8】



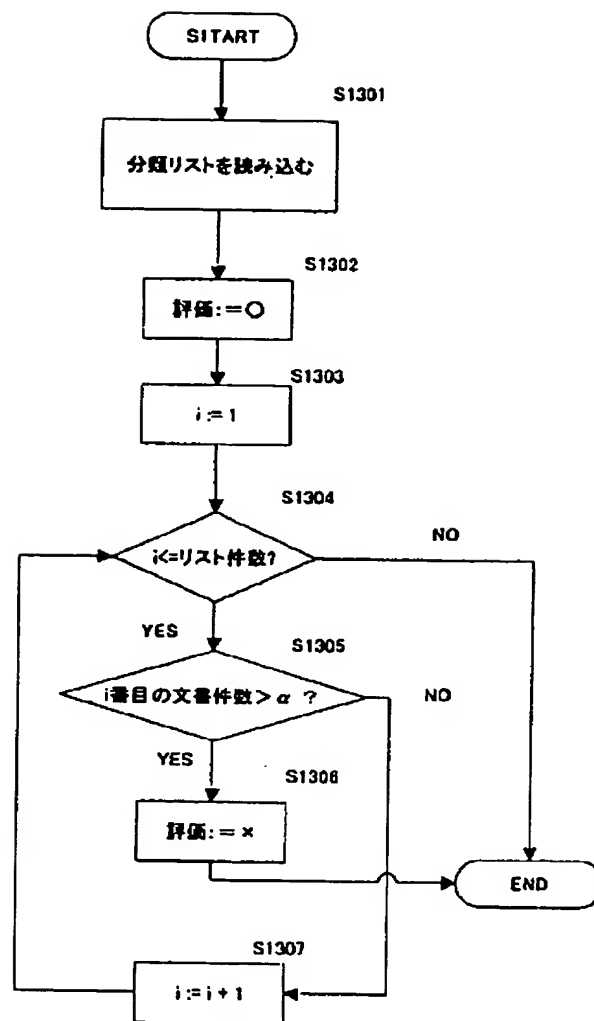
【図 10】



【図11】



【図13】



フロントページの続き

(51) Int. Cl. 7

識別記号

F I  
G 0 6 F 15/401

テーマコード (参考)  
3 1 0 D